Solution 1: Entropy

(a) Let a, b, x, and y denote the realisations of the random variables A, B, X, and Y, respectively. Each event (a, b) is associated with exactly one event (x, y) and the probability for such an event is given by

$$p_{AB}(a,b) = p_{XY}(x,y) = \frac{1}{6} \cdot \frac{1}{2} = \frac{1}{12}$$

Consequently, we obtain for the joint entropy

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p_{X,Y}(x,y) \log_2 p_{XY}(x,y) = -12 \cdot \frac{1}{12} \log_2 \frac{1}{12} \\ &= \log_2 12 \\ &= 2 + \log_2 3 \end{aligned}$$

Below we list the possible values of the random variables X and Y, the associated events (a, b), and the probability masses $p_X(x)$ and $p_Y(y)$.

x	events (a,b)	$p_X(x)$
1	$(1,0)$	1/12
2	$(2,0),(1,1)$	1/6
3	$(3,0),(2,1)$	1/6
4	$(4,0),(3,1)$	1/6
5	$(5,0),(4,1)$	1/6
6	$(6,0),(5,1)$	1/6
7	$(6,1)$	1/12

y	events (a,b)	$p_Y(y)$
0	$(1,1)$	1/12
1	$(1,0),(2,1)$	1/6
2	$(2,0),(3,1)$	1/6
3	$(3,0),(4,1)$	1/6
4	$(4,0),(5,1)$	1/6
5	$(5,0),(6,1)$	1/6
6	$(6,0)$	1/12

The random variable $X = A + B$ can take the values 1 to 7. The probability masses $p_X(x)$ for the values 1 and 7 are equal to 1/12, since they correspond to exactly one event. The probability masses for the values 2 to 6 are equal to 1/6, since each of these values corresponds to two events (a, b). An analogue result is obtained for the random variable $Y = A - B$.

The marginal entropies are given by

$$\begin{aligned} H(X) &= -\sum_{x} p_X(x) \log_2 p_X(x) \\ &= -2 \cdot \frac{1}{12} \log_2 \frac{1}{12} - 5 \cdot \frac{1}{6} \log_2 \frac{1}{6} \\ &= \frac{1}{6} \cdot (\log_2 4 + \log_2 3) + \frac{5}{6} \cdot (\log_2 2 + \log_2 3) \\ &= \frac{7}{6} + \log_2 3 \end{aligned}$$

and for Y

$$
\begin{aligned}
H(Y) &= -\sum_y p_Y(y)\log_2 p_Y(y) \\
&= -2\cdot\frac{1}{12}\log_2\frac{1}{12} - 5\cdot\frac{1}{6}\log_2\frac{1}{6} \\
&= \frac{1}{6}\cdot(\log_2 4 + \log_2 3) + \frac{5}{6}\cdot(\log_2 2 + \log_2 3) \\
&= \frac{7}{6} + \log_2 3
\end{aligned}
$$

We can determine the conditional entropies using

$$H(X|Y) = H(X,Y) - H(Y) = 2 + \log_2 3 - \frac{7}{6} - \log_2 3 = \frac{5}{6}$$

$$H(Y|X) = H(X,Y) - H(X) = 2 + \log_2 3 - \frac{7}{6} - \log_2 3 = \frac{5}{6}$$

The mutual information $I(X;Y)$ can be determined acording to

$$I(X;Y) = H(X) - H(X|Y) = \frac{7}{6} + \log_2 3 - \frac{5}{6} = \frac{1}{3} + \log_2 3$$

or

$$I(X;Y) = H(Y) - H(Y|X) = \frac{7}{6} + \log_2 3 - \frac{5}{6} = \frac{1}{3} + \log_2 3$$

(b) Using the definition of mutual information for discrete random variables, $I(X;Y) = H(Y) - H(Y|X)$, we can write

$$
\begin{aligned}
I(X;X+Y) - I(Y;X+Y) &= H(X+Y) - H(X+Y|X) - H(X+Y) + H(X+Y|Y) \\
&= H(X|Y) - H(Y|X) \\
&= H(X) - H(Y)\,.
\end{aligned}
$$

The first step follows from the fact that modifying the mean of a pmf doesn't change the entropy. For the second step, we used the fact that the conditional entropy $H(X|Y)$ is equal to the marginal entropy $H(X)$ for independent random variables X and Y.

Solution 2: The Mutual Information of Three Variables

(a) According to the definition of mutual information, we have

$$
\begin{aligned}
&I(X;Y) - I(X;Y|Z) \\
&= \sum_x\sum_y p(x,y)\log\frac{p(x,y)}{p(x)p(y)} - \underbrace{\sum_z\sum_x\sum_y p(z)p(x,y|z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)}}_{\text{The definition of conditional mutual information}} \\
&= \sum_x\sum_y\sum_z p(x,y,z)\log\frac{p(x,y)}{p(x)p(y)} - \sum_x\sum_y\sum_z p(x,y,z)\log\frac{p(x,y|z)p(z)^2}{p(x|z)p(y|z)p(z)^2} \\
&= \sum_x\sum_y\sum_z p(x,y,z)\log\frac{p(x,y)}{p(x)p(y)} - \sum_x\sum_y\sum_z p(x,y,z)\log\frac{p(x,y,z)p(z)}{p(x,z)p(y,z)} \\
&= \sum_x\sum_y\sum_z p(x,y,z)\log\left(\frac{p(x,y)p(x,z)p(y,z)}{p(x)p(y)p(z)p(x,y,z)}\right) \\
&= I(X;Y;Z).
\end{aligned}
\tag{1}
$$

(b) Using the lemma we just proved, we obtain:

$$\begin{aligned}
&I(X;Y|Z) + I(Y;Z) - I(Y;Z|X) \\
&= I(X;Y) - I(X;Y;Z) + I(Y;Z) - I(Y;Z) + I(X;Y;Z) \\
&= I(X;Y).
\end{aligned} \tag{2}$$

(P.S., a recent paper [1] provides a good example of how this relation is used in the research of explainability.)

Solution 3: Smoothed Cross-Entropy Loss

(a) The empirical risk is

$$\begin{aligned}
R_{\text{emp}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{g}\tilde{d}_k^{(i)}\log\left(\frac{\tilde{d}_k^{(i)}}{\pi_k(\mathbf{x}^{(i)}|\theta)}\right)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\sum_{k=1}^{g}\tilde{d}_k^{(i)}\log\tilde{d}_k^{(i)} - \tilde{d}_k^{(i)}\log\pi_k(\mathbf{x}^{(i)}|\theta)\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{g}\tilde{d}_k^{(i)}\log\pi_k(\mathbf{x}^{(i)}|\theta) + Const.
\end{aligned} \tag{3}$$

(b) The smoothed cross-entropy is implemented as follows:

```r
#' @param label ground truth vector of the form (n_samples,).
#'  Labels should be "1","2","3" and so on.
#' @param pred Predicted probabilities of the form (n_samples,n_labels)
#' @param smoothing Hyperparameter for label-smoothing

smoothed_ce_loss <- function(
label,
pred,
smoothing){

  num_samples <- NROW(pred)
  num_classes<- NCOL(pred)

  # Let's make some assertions:
  #  label should be a 1-D array.one-hot encoded label is not necessary
  stopifnot(NCOL(label)==1)
  # smoothing hyperparameter in allowed range
  stopifnot((smoothing>=0 & smoothing <= 1))
  # Same amount of rows in labels and predictions
  stopifnot((NROW(label)== num_samples))
  # Predicted probabilities must have as many columns as labels
  stopifnot(length(unique(label)) == num_classes)

  #Calculate the base level
  smoothing_per_class <- smoothing / num_classes

  # build the label matrix. Shape = [ num_samples, num_classes]
  # Start with the base level
  smoothed_labels_matrix = matrix(smoothing_per_class,
                                  nrow=num_samples,ncol=num_classes)
  # Add the smoothed correct labels
  true_labels_loc=cbind(1:num_samples, label)
  smoothed_labels_matrix[true_labels_loc]= 1 - smoothing + smoothing_per_class
  cat("Labels matrix:\n")
  print(smoothed_labels_matrix)
```

```
  # Calculate the loss
  cat("Loss for each sample:\n ",
      rowSums(- smoothed_labels_matrix * log(pred)))

  loss <- mean(rowSums(- smoothed_labels_matrix * log(pred)))
  cat("\n Loss:\n",loss)

  return (loss)
}
```

```
  # Let's build a "confident model", the model has very high predicted
  #probabilities for one of the labels
  label= c(1,2,2,3,1)
  pred= rbind(
          c(0.85,0.10,0.05),
          c(0.05,0.9,0.05),
          c(0.02,0.95,0.03),
          c(0.13,0.02,0.85),
          c(0.86,0.04,0.1))

  # cross entropy means smoothing=0
  smoothing=0
  loss<-smoothed_ce_loss(label,pred,smoothing)

## Labels matrix:
##      [,1] [,2] [,3]
## [1,]    1    0    0
## [2,]    0    1    0
## [3,]    0    1    0
## [4,]    0    0    1
## [5,]    1    0    0
## Loss for each sample:
##   0.1625189 0.1053605 0.05129329 0.1625189 0.1508229
##  Loss:
##  0.1265029

  # Smoothed cross entropy
  smoothing=0.2
  loss_smooth<-smoothed_ce_loss(label,pred,smoothing)

## Labels matrix:
##            [,1]       [,2]       [,3]
## [1,] 0.86666667 0.06666667 0.06666667
## [2,] 0.06666667 0.86666667 0.06666667
## [3,] 0.06666667 0.86666667 0.06666667
## [4,] 0.06666667 0.06666667 0.86666667
## [5,] 0.86666667 0.06666667 0.06666667
## Loss for each sample:
##   0.4940709 0.4907434 0.5390262 0.537666 0.4988106
##  Loss:
##  0.5120634
```

References

[1] Rong, Yao, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. "A consistent and efficient evaluation strategy for attribution methods." In International Conference on Machine Learning, pp. 18770-18795.

PMLR, 2022.